Pattern-based Aggregation of Named Entity Extractors

Kofi Boakye

Collaborators: Tracy Lemmond, Paul Kidwell, Nathan Perry, Joe Guensche, William Hanley, Ryan Prenger, and Ron Glaser

National Security Engineering Division

We present an inference framework that leverages the joint characteristics of entity extractor error processes via a pattern-based representation of entity data. This approach has been shown to produce statistically significant improvements in entity extraction and to mitigate the weak performance of entity extractors operating under suboptimal conditions. Moreover, this aggregation methodology provides a framework for quantifying uncertainty in extracted entity output, and it can readily adapt to sparse data conditions.

Motivation

Entity extraction
Detecting people, locations, etc. in text documents

Microsoft and Yahoo announced their search partnership this morning. In a nutshell, Microsoft will power search technology for both companies, while Yahoo will take over ad sales. Here are my notes from this morning's Yahoo-Microsoft conference call:

Carol Bartz: This is a great day for Yahoo. A game changer. benefits for Yahoo. half of all Internet users come to us, but face a formidable competitor in search.

- Supports search engines, knowledge discovery
- Improved performance possible by aggregating output of multiple extractors

Aggregation: The General Approach

Step 1: Construct a meta-entity

Nikko Securities Co. New Brunswick Co.	*BALIE*
Nikko Securities Co New Brunswick Scientific Co	*GATE*
Nikko Securities Co. New Brunswick Scientific Co.	*LingPipe*
Nikko Securities Co. New Brunswick Scientific Co.	*SNER*

"Nikko Securities Co. New Brunswick Scientific Co." ⇐ **Meta-entity**

Step 2: Construct a hypothesis space

- *Nikko Securities Co.*
- *Nikko Co. New*
- *Nikko Securities Co. New Brunswick*
- *Nikko New*

Step 3: Assign probabilities to hypotheses

Step 4: Rank hypotheses

Nikko Securities Co. New Brunswick Scientific Co.	*0.42*
Nikko Securities Co. Scientific Co.	*0.33*
Nikko Securities Co.	*0.06*

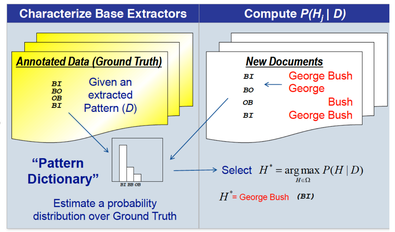
The Pattern Meta-Entity Extractor (PME)

Implicitly model errors to compute $P(H_j \mid D)$

"… of Nikko Securities Co. New Brunswick Scientific Co., a maker of…"

Nikko Securities Co. New Brunswick Scientific Co. *BIBIII*	*Truth*
Nikko Securities Co. New Brunswick Co. *BBBIOB*	*BALIE*
Nikko Securities Co New Brunswick Scientific Co *BIBBII*	*GATE*
Nikko Securities Co. New Brunswick Scientific Co. *BIBIBI*	*LingPipe*
Nikko Securities Co. New Brunswick Scientific Co. *BIIIII*	*SNER*

Encode structural patterns in the extracted data



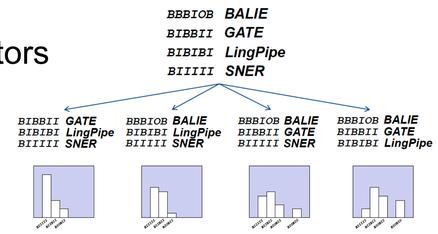
Extensions: Unprecedented Patterns

"Stepping Down"

Aggregate over subsets of extractors

Two approaches:

1) Simple k-way decision
   - Step down until found
2) Lower bound maximization
   - Step down to "best" combination, based on lower Bayesian bound



Sequential Meta-Entity Model

Decompose joint probability of pattern by columns and apply n-gram model:

$$P(d_m, H_{mj}) = P(c_1)\prod_{t=2}^{s} P(c_t \mid c_{t-1},\ldots,c_{t-n})$$

Choose hypothesis maximizing joint probability

	c_1	c_2	c_3	c_4
d_{m1}	2	1	2	1
d_{m2}	2	1	0	2
d_{m3}	2	0	0	2
H_{mj}	2	1	0	2

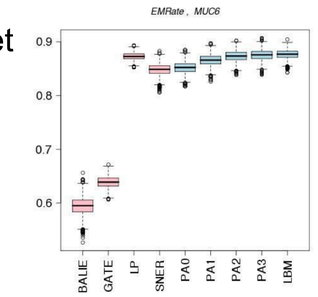
Results

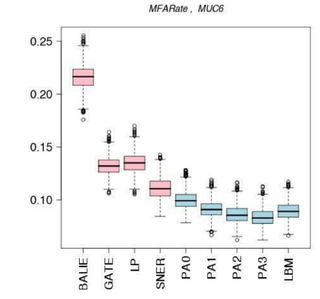
Data: MUC6 evaluation set
- Wall Street Journal text
- ~7600 entities

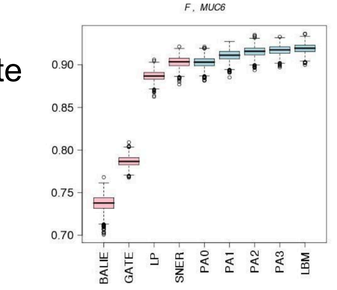
Extractors:
1) BALIE – Unsupervised learning
2) GATE – Rule-based
3) LingPipe – Hidden Markov Models
4) SNER – Conditional Random Fields

Metrics:
1) Exact match rate
2) Miss + False alarm rate
3) F-measure

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$







This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-485133